SUPPLEMENTARY TEXT: EFFICIENT COMPUTATION WITH A LINEAR MIXED MODEL ON LARGE-SCALE DATA SETS WITH APPLICATIONS TO GENETIC STUDIES

By Matti Pirinen, Peter Donnelly and Chris C.A. Spencer

University of Oxford

In this supplement to "Efficient Computation with a Linear Mixed Model on Large-scale Data Sets with Applications to Genetic Studies" we give the details of the application of the linear mixed model to binary data, of the conditional maximization of the likelihood function and of the Bayesian computations.

Throughout this text we consider the linear mixed model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varrho} + \boldsymbol{\varepsilon}, \tag{0.1}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is the vector of responses on n subjects, $\boldsymbol{X} = (x_{ik})$ is the $n \times K$ matrix of predictor values on the subjects, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ collects the (unknown) linear effects of the predictors on the responses $\boldsymbol{Y}$ and random effects $\boldsymbol{\varrho}$ and $\boldsymbol{\varepsilon}$ are assigned distributions

$$\boldsymbol{\varrho}|(\eta, \sigma^2) \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{R}) \text{ and } \boldsymbol{\varepsilon}|(\eta, \sigma^2) \sim \mathcal{N}(0, (1-\eta)\sigma^2\boldsymbol{I}), \tag{0.2}$$

where $\boldsymbol{R}$ is a known positive semi-definite $n \times n$ matrix, $\boldsymbol{I}$ is the $n \times n$ identity matrix, and parameters $\sigma^2 > 0$ and $\eta \in [0, 1]$ determine how the variance is divided between $\boldsymbol{\varrho}$ and $\boldsymbol{\varepsilon}$.

1. Binary data. For 0-1 valued responses $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, a logistic regression model assumes that

$$p_i = P(y_i = 1|\boldsymbol{X}, \alpha, \boldsymbol{\gamma}) = \frac{\exp(\alpha + \boldsymbol{X}_i\boldsymbol{\gamma})}{1 + \exp(\alpha + \boldsymbol{X}_i\boldsymbol{\gamma})},$$

where the row i of $\boldsymbol{X}$ is denoted by $\boldsymbol{X}_i$, the effects of the predictors are in vector $\boldsymbol{\gamma}$ and α is the population base-line effect. (Note that here the base-line effect has been explicitly separated from the $\boldsymbol{X}$ matrix.) The log-likelihood function for exchangeable observations is

$$L_b(\alpha, \boldsymbol{\gamma}) = \log P(\boldsymbol{Y}|\alpha, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(y_i \log p_i + (1 - y_i) \log(1 - p_i)\right),$$

where we use the subscript b to denote "binary".

First order approximation (FOA). By treating α and $\boldsymbol{\gamma}$ as known, by mean-centring the predictors and by expanding p_i as a Taylor series around the mean, $\boldsymbol{X}_i = 0$, we have

$$p_i \approx \frac{e^\alpha}{1+e^\alpha} + \frac{e^\alpha}{(1+e^\alpha)^2}\boldsymbol{X}_i\boldsymbol{\gamma} + \frac{e^\alpha(1-e^\alpha)}{2(1+e^\alpha)^3}(\boldsymbol{X}_i\boldsymbol{\gamma})^2 + \dots. \tag{1.1}$$

When the effects are small on the log-odds scale in the sense that $|\boldsymbol{X}_i\boldsymbol{\gamma}|$ is small, then $(\boldsymbol{X}_i\boldsymbol{\gamma})^2 \approx 0$ and the probability p_i is accurately approximated by a linear function of the predictors $p_i \approx \mu + \boldsymbol{X}_i\boldsymbol{\beta}$ constrained to lie in $[0,1]$. According to (1.1), the parameters are transformed between logistic $(\alpha, \boldsymbol{\gamma})$ and linear $(\mu, \boldsymbol{\beta})$ scales as

$$\begin{array}{llll} \alpha = \log\left(\frac{\mu}{1-\mu}\right), & \gamma_k = \frac{\beta_k}{\mu(1-\mu)}, & \text{for } k = 1,\dots,K, \\ \mu = \frac{e^\alpha}{1+e^\alpha}, & \beta_k = \gamma_k \frac{e^\alpha}{(1+e^\alpha)^2}, & \text{for } k = 1,\dots,K. \end{array} \tag{1.2}$$

The score and the Hessian of the logistic regression model are

$$\frac{\partial L_b}{\partial \boldsymbol{\gamma}} = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{p}) \tag{1.3}$$

$$\frac{\partial^2 L_b}{\partial \boldsymbol{\gamma}^2} = -\boldsymbol{X}^T \mathrm{diag}(p_i(1-p_i))\boldsymbol{X}, \tag{1.4}$$

where we have included the base-line parameter α in $\boldsymbol{\gamma}$ and augmented $\boldsymbol{X}$ accordingly with a column of ones, and $\boldsymbol{p} = (p_1, \dots, p_n)^T$ is a function of $\boldsymbol{\gamma}$. By using the small-effect approximation $\boldsymbol{p} \approx \boldsymbol{X}\boldsymbol{\beta}$ as in the derivation of (1.2), but now with μ included in $\boldsymbol{\beta}$, we see that the score (1.3) is approximately zero at the least squares estimate $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Thus, if the assumption of small effects is valid, an application of the least squares method (i.e. the maximum likelihood (ML) estimation in the standard linear model) to binary data and the transformation of the parameters to the log-odds scale using (1.2) should give a good approximation to the ML estimates of the logistic regression model. The sampling variance of the coefficients could be approximated by the inverse of the negative Hessian (1.4) at the estimated maximum or by transformation (1.6) explained below. Despite the simplicity of this linear approximation, we are not aware of its previous formal derivation, although similar ideas have been applied before, for example by Denby, Kafadar and Land (1998). From now on we call it the *first order approximation* (FOA) to distinguish it from a more accurate approximation that we have established particularly for our genetics application.

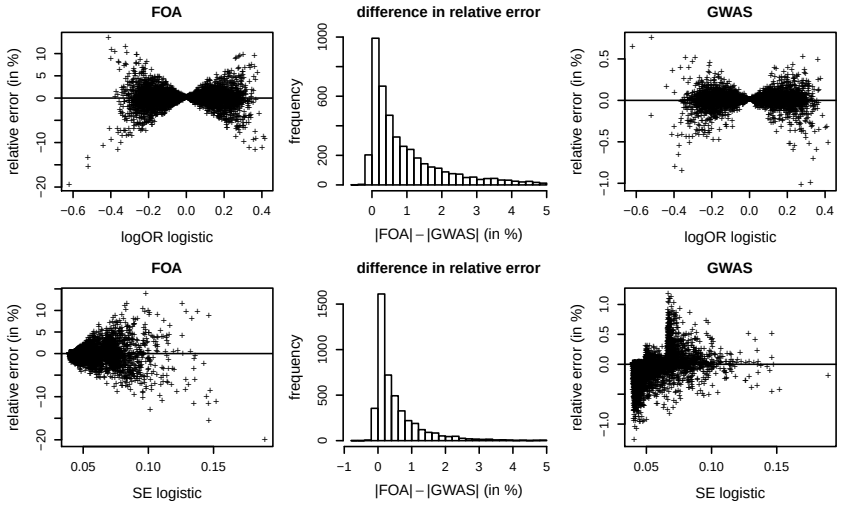

FIG 1. *Comparing the first order approximation (FOA) and the GWAS approximation. Panels include 4,500 binary variants considered in Figure 2 of the main text. The leftmost column shows relative errors (in percentages) between the FOA and the maximum likelihood estimates from the logistic regression model as a function of the estimated log-odds ratios (top) and standard errors (bottom). The rightmost column shows similar results for the GWAS approximation. The middle column shows histograms of the differences between absolute values of the relative errors (in percentages) from the FOA and the GWAS approximation, truncated from above at 5%. logOR, log-odds ratio; SE, standard error.*

GWAS approximation. We consider a GWAS setting in which the case-control status is regressed on the population mean and the reference allele count. By examining the second and third order terms of series (1.1) and carrying out some empirical testing we found that the relative differences between the log-odds estimates from the FOA and the ML estimates from the logistic regression model are accurately described by

$$(1.5)\quad r(\bar{\gamma},\theta,\phi) = 0.5(1-2\phi)(1-2\theta)\bar{\gamma} - (0.084 + 0.9\phi(1-2\phi)\theta(1-\theta))\bar{\gamma}^2,$$

where θ is the frequency of the reference allele, $\bar{\gamma}$ is the log-odds estimate for the reference allele from the FOA and ϕ is the proportion of cases in the data. This can be used for adjusting both the estimates: $\widehat{\gamma} = \bar{\gamma}/(1 + r(\bar{\gamma},\theta,\phi))$, and their standard errors. Figure 1 above shows the improvement of this GWAS approximation over the FOA.

Mixed model. When we model the binary responses $\boldsymbol{Y}$ as correlated according to the variance structure $\sigma^2\boldsymbol{\Sigma}$ where the matrix $\boldsymbol{\Sigma}$ is known, an analogous estimate of the parameters is the generalized least squares (GLS) solution $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$ which can be transformed to

the log-odds scale using (1.2) (and possibly adjusting by the GWAS approximation). In this case the sampling variance on the linear scale can be approximated using the GLS estimate $\widehat{\boldsymbol{V}}_\beta = \widehat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$, where $\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$. The corresponding estimates on the log-odds scale are then given by the delta-method:

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}} = \boldsymbol{J}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{J}^T, \text{ where } \boldsymbol{J}_{ij} = \left(\frac{\partial \gamma_i}{\partial \beta_j}\right)_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}. \tag{1.6}$$

We note that in the most general case of our linear mixed model (0.1) the covariance structure $\boldsymbol{\Sigma} = \eta\boldsymbol{R} + (1-\eta)\boldsymbol{I}$ includes the parameter η. If η is estimated by maximum likelihood we cannot any more justify the method as a pure least squares method. In any case the empirical results in the main text suggest that the procedure works well in our application.

The standard way of finding ML estimates for logistic regression is known as iteratively reweighted least squares (Nelder and Wedderburn, 1972). As an instance of the Newton-Raphson algorithm it is based on the second order Taylor series approximation of the log-likelihood and results in a series of least squares problems where the outcome variable and the diagonal covariance matrix vary between each iteration. In contrast, our first order approximation is based on the linear approximation of the probabilities p_i (not the log-likelihood) and is available after a single application of the least squares method to the original binary data, but with a downside that it is accurate only in the case of small effect sizes.

Equivalence between the trend test and the linear model. Above we showed how the linear model can estimate the effects on the log-odds scale. Next we give another justification for the application of the linear model to case-control data by showing that for large sample sizes the likelihood ratio test for the SNP effect in the standard linear model is equivalent to the Armitage trend test of the genotype counts (Armitage, 1955). The trend test is widely-used for analysing case-control GWAS and in this context is also equivalent to a score test of a logistic regression model. Previously, connections between the trend test and the linear model in the GWAS context have been discussed by Kang et al. (2010); also Astle and Balding (2009) give conditions under which the linear model can be applied to case-control data.

Suppose that we have genotype data on S cases and R controls with $n = S + R$, and denote the mean-centred genotype of individual i by x_i and the binary phenotype by $y_i \in \{0, 1\}$. The trend test-statistic can be written

as T^2/V where

$$\begin{aligned} T &= \frac{1}{S}\sum_{i\in S} x_i - \frac{1}{R}\sum_{i\in R} x_i = \frac{n}{SR}\sum_{i\in S} x_i \\ V &= \frac{1}{SR}\sum_{i=1}^{n} x_i^2, \end{aligned}$$

and it has an asymptotic χ^2_1-distribution under the null hypothesis of no linear trend in the genotype frequencies between the cases and the controls (Astle and Balding, 2009).

The maximum likelihood estimates for the linear models $M_0 : y_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $M_1 : y_i \sim \mathcal{N}(\mu_1 + \beta_1 x_i, \sigma_1^2)$ are

$$\begin{aligned} \hat{\mu}_0 &= \phi, \\ \hat{\mu}_1 &= \phi, \\ \hat{\beta}_1 &= \frac{\sum_{i\in S} x_i}{\sum_{i=1}^n x_i^2}, \\ \widehat{\sigma_0^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_0)^2 = \phi(1-\phi), \\ \widehat{\sigma_1^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_1 - \hat{\beta}_1 x_i)^2 = \phi(1-\phi) - \frac{(\sum_{i\in S} x_i)^2}{n(\sum_{i=1}^n x_i^2)}, \end{aligned}$$

where $\phi = S/n$. The likelihood ratio statistic is

$$\begin{aligned} 2\log\left(\frac{L_1(\hat{\mu}_1, \hat{\beta}_1, \widehat{\sigma_1^2})}{L_0(\hat{\mu}_0, \widehat{\sigma_0^2})}\right) &= n\log\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma_1^2}}\right) \\ &= -\log\left(\left(1 - \frac{(\sum_{i\in S} x_i)^2}{n\phi(1-\phi)\sum_{i=1}^n x_i^2}\right)^n\right) \\ &\xrightarrow[n\to\infty]{} -\log\left(\exp\left(-\frac{(\sum_{i\in S} x_i)^2}{\phi(1-\phi)\sum_{i=1}^n x_i^2}\right)\right) \\ &= \frac{(\sum_{i\in S} x_i)^2}{\phi(1-\phi)\sum_{i=1}^n x_i^2} = T^2/V. \end{aligned}$$

Here the convergence is derived from a basic property of the exponential function: $(1+a/n)^n \to e^a$ as $n \to \infty$, for any real value a. Thus the likelihood ratio statistics approaches the trend test statistic as $n \to \infty$.

2. Likelihood analysis. The log-likelihood function for model (0.1) is

$$L(\boldsymbol{\beta}, \eta, \sigma^2) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\Sigma} = \eta\boldsymbol{R} + (1-\eta)\boldsymbol{I}$, $c = -\frac{n}{2}\log(2\pi)$ and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

The eigenvalue decomposition of the positive semi-definite matrix $\boldsymbol{R}$ yields an orthonormal $n \times n$-matrix $\boldsymbol{U}$ of eigenvectors and a diagonal $n \times n$-matrix $\boldsymbol{D}$ of non-negative eigenvalues for which $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ (see e.g. Golub and Van Loan (1996)). Because $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ (orthonormality) it follows that

$$\begin{aligned}
\boldsymbol{\Sigma} &= \eta\boldsymbol{R} + (1-\eta)\boldsymbol{I} \\
&= \eta\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T + (1-\eta)\boldsymbol{U}\boldsymbol{I}\boldsymbol{U}^T \\
&= \boldsymbol{U}(\eta\boldsymbol{D} + (1-\eta)\boldsymbol{I})\boldsymbol{U}^T, \\
\boldsymbol{\Sigma}^{-1} &= \boldsymbol{U}(\eta\boldsymbol{D} + (1-\eta)\boldsymbol{I})^{-1}\boldsymbol{U}^T, \\
|\boldsymbol{\Sigma}| &= \prod_{i=1}^{n}(1 + \eta(d_i - 1)),
\end{aligned}$$

where d_i is the ith eigenvalue of $\boldsymbol{R}$, that is, the element (i,i) of $\boldsymbol{D}$. The inverse $\boldsymbol{\Sigma}^{-1}$ is defined for all $\eta \in [0,1]$ unless some d_i is zero in which case we restrict the model to the values $\eta < 1$.

By transformations $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}^T\boldsymbol{Y}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{U}^T\boldsymbol{X}$, and $\widetilde{\boldsymbol{\Sigma}} = \eta\boldsymbol{D} + (1-\eta)\boldsymbol{I}$ the log-likelihood becomes

$$\begin{aligned}
& L(\boldsymbol{\beta}, \eta, \sigma^2) \\
= \; & c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|\widetilde{\boldsymbol{\Sigma}}|) - \frac{1}{2\sigma^2}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}) \\
= \; & c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(1 + \eta(d_i - 1)) - \sum_{i=1}^{n}\frac{([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{2\sigma^2(1 + \eta(d_i - 1))},
\end{aligned}$$

where $[\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i$ is the ith element of vector $\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}$. For each set of values of the parameters, the evaluation of the log-likelihood requires $\mathcal{O}(nK)$ basic operations, where n is the number of individuals (rows of $\boldsymbol{X}$ matrix) and K is the number of predictors in the model (columns of $\boldsymbol{X}$).

2.1. *Conditional maximization.* To maximize the log-likelihood we use a standard optimization technique of conditional maximization. After initializing the parameters to values $(\boldsymbol{\beta}^{(0)}, \eta^{(0)}, (\sigma^2)^{(0)})$, we iterate the following

three step process until convergence:

$$\begin{aligned}\boldsymbol{\beta}^{(j+1)} &= \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \eta^{(j)}, (\sigma^2)^{(j)}) \\ (\sigma^2)^{(j+1)} &= \arg\max_{\sigma^2} L(\boldsymbol{\beta}^{(j+1)}, \eta^{(j)}, \sigma^2) \\ \eta^{(j+1)} &= \arg\max_{\eta} L(\boldsymbol{\beta}^{(j+1)}, \eta, (\sigma^2)^{(j+1)}),\end{aligned}$$

where the superscripts in parentheses denote the iteration. We have not established theoretical conditions which would guarantee that the process finds the global maximum, but we know that conditional on η we always find the global maximum with respect to $\boldsymbol{\beta}$ and σ^2. Furthermore, in the comparisons with the EMMA algorithm we have not found a single data set where the algorithm would have failed (see the main text). If such exist in some applications, then one could run the algorithm several times starting from different initial values. Steps 1 and 2 are done analytically by using standard results on linear models, and step 3 is done by numerical maximization using some ideas from Kang et al. (2008).

Step 1: The derivative

$$\frac{\partial L(\boldsymbol{\beta}, \eta, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{X}}$$

is zero at $\widehat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{Y}}$ assuming that matrix $\widetilde{\boldsymbol{X}}$ is of full column rank. Under the same assumption the Hessian $\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2}(\widehat{\boldsymbol{\beta}}) = -\frac{1}{\sigma^2} \widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{X}}$ is negative definite and thus the function $\boldsymbol{\beta} \to L(\boldsymbol{\beta}, \eta, \sigma^2)$ attains its global maximum at $\widehat{\boldsymbol{\beta}}$.

Step 2: The derivative

$$\frac{\partial L(\boldsymbol{\beta}, \eta, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \frac{([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{(1 + \eta(d_i - 1))}$$

is zero at $\widehat{\sigma^2} = \frac{A}{n}$, where $A = \sum_{i=1}^{n} \frac{([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{(1+\eta(d_i-1))}$. The second derivative $\frac{\partial^2 L}{\partial (\sigma^2)^2}(\widehat{\sigma^2}) = \frac{-n^3}{2A^2} < 0$ thus showing that the function $\sigma^2 \to L(\boldsymbol{\beta}, \eta, \sigma^2)$ attains its global maximum at $\widehat{\sigma^2}$. (Actually, since the value $\widehat{\boldsymbol{\beta}}$ in step 1 does not depend on σ^2, the steps 1 and 2 together give the global maximum of the function $(\boldsymbol{\beta}, \sigma^2) \to L(\boldsymbol{\beta}, \eta, \sigma^2)$.)

Step 3: We use a Newton-Raphson method to find zeros of the derivative

$$\frac{\partial L(\boldsymbol{\beta}, \eta, \sigma^2)}{\partial \eta} = \frac{1}{2} \sum_{i=1}^{n} \frac{d_i - 1}{1 + \eta(d_i - 1)} \left(\frac{([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{\sigma^2(1 + \eta(d_i - 1))} - 1 \right).$$

We divide the interval $[0,1]$ into m subintervals by points $\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$ and evaluate the derivative at each of these points. If the sign of the derivative changes from positive to negative within an interval, we apply the Newton-Raphson algorithm to find a zero within that interval. Finally we choose the maximum of the log-likelihood values among the local maxima (zeros of the derivative) or the values at the endpoints. A problem would occur if there were several zeros within a single interval because this algorithm would find at most only one of them. To reduce chances of such an event one should in principle use a relatively large number of subintervals m. In our examples, we have used $m = 10$.

2.2. *The second derivatives.* Asymptotic likelihood theory allows us to estimate the standard errors of the parameters by using the inverse of the observed information matrix $\mathcal{I}$ at the MLE. The elements of $\mathcal{I}$ are

$$\mathcal{I}_{ij} = -\frac{\partial^2 L}{\partial\theta_i \partial\theta_j}(\widehat{\boldsymbol{\theta}}),$$

where $(\theta_1, \ldots, \theta_{K+2}) = (\beta_1, \ldots, \beta_K, \eta, \sigma^2)$. Straightforward calculations show that the second derivatives are

$$\begin{aligned}
\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2} &= -\frac{1}{\sigma^2}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}} \\
\frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial(\sigma^2)} &= -\frac{1}{\sigma^4}(\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta})^T\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}} \\
\frac{\partial^2 L}{\partial \beta_k \partial \eta} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(d_i-1)[\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i\widetilde{\boldsymbol{X}}_{ik}}{(1+\eta(d_i-1))^2} \\
\frac{\partial^2 L}{\partial(\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}\frac{([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{1+\eta(d_i-1)} \\
\frac{\partial^2 L}{\partial(\sigma^2)\partial\eta} &= -\frac{1}{2\sigma^4}\sum_{i=1}^{n}\frac{(d_i-1)([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{(1+\eta(d_i-1))^2} \\
\frac{\partial^2 L}{\partial\eta^2} &= \frac{1}{2}\sum_{i=1}^{n}\left(\frac{d_i-1}{1+\eta(d_i-1)}\right)^2\left(1 - \frac{2([\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}]_i)^2}{\sigma^2(1+\eta(d_i-1))}\right).
\end{aligned}$$

3. Bayesian computation. In a Bayesian version of the mixed model, we combine the sampling distribution

$$\boldsymbol{Y}|(\boldsymbol{\beta}, \sigma^2, \eta) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \eta\sigma^2\boldsymbol{R} + (1-\eta)\sigma^2\boldsymbol{I}),$$

with the prior distributions

$$\begin{aligned}(\boldsymbol{\beta}, \sigma^2) &\sim \text{Normal-Inverse-Gamma}(\boldsymbol{m}, \boldsymbol{V}, a, b),\\ \eta &\sim \text{Beta}(r, t),\end{aligned}$$

where $a, b, r, t > 0$ are scalar parameters, $\boldsymbol{m}$ is a K-dimensional vector and $\boldsymbol{V}$ is a $K \times K$-matrix. These choices of prior distributions lead to the following marginal properties:

$$\begin{array}{lll}\boldsymbol{\beta} \sim t_{2a}(\boldsymbol{m}, 2b\boldsymbol{V}) & \mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{m} & \mathrm{Var}(\boldsymbol{\beta}) = \frac{b}{a-1}\boldsymbol{V}\\ \sigma^2 \sim \text{Inv-Gamma}(a, b) & \mathrm{E}(\sigma^2) = \frac{b}{a-1} & \mathrm{Var}(\sigma^2) = \frac{b^2}{(a-1)^2(a-2)}\\ \eta \sim \text{Beta}(r, t) & \mathrm{E}(\eta) = \frac{r}{r+t} & \mathrm{Var}(\eta) = \frac{rt}{(r+t)^2(r+t+1)}.\end{array}$$

An intuitive description of the Normal-Inverse-Gamma (NIG) distribution is that a pair $(\boldsymbol{\beta}, \sigma^2)$ is generated from $\text{NIG}(\boldsymbol{m}, \boldsymbol{V}, a, b)$ by first sampling $\sigma^2 \sim \text{Inverse-Gamma}(a, b)$ and then $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(\boldsymbol{m}, \sigma^2\boldsymbol{V})$.

The most notable restriction of these priors is that $\boldsymbol{\beta}$ and σ^2 are a priori dependent (see O'Hagan and Forster (2004)). To adjust the prior in a particular setting it is often helpful to standardize both the quantitative responses and each continuous predictor. The GWAS software SNPTEST2 uses a similar prior distribution for analyzing quantitative traits with the standard linear model and some guidelines for prior specification can be found in its manual[1].

The steps to carry out analytic integration of $\boldsymbol{\beta}$ and σ^2 in the mixed model considered here have the same form as the corresponding steps in the general linear model (O'Hagan and Forster, 2004). This is an advantage of our parameterization compared to a previous treatment of this mixed model by Sorensen and Gianola (2002); for details of the differences, see the discussion at the end of this section. Another novelty of our work is to show that the marginal likelihood computations can be done efficiently using the same matrix decomposition that was introduced for ML estimation in the previous section. This is crucial in order that a large number of $\boldsymbol{X}$ matrices can be analyzed efficiently in the Bayesian framework. To our knowledge, this topic has not previously been considered in the literature.

3.1. *Computation.* As before, the likelihood part of the model is

$$p(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, \eta) = (2\pi)^{-\frac{n}{2}}|\sigma^2\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right)$$

[1] www.stats.ox.ac.uk/~marchini/software/gwas/snptest

with $\boldsymbol{\Sigma} = \eta \boldsymbol{R} + (1-\eta)\boldsymbol{I}$ and the prior density $p(\boldsymbol{\beta}, \sigma^2, \eta) = p(\eta)p(\boldsymbol{\beta}, \sigma^2)$ is composed of

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2) &= \frac{b^a (\sigma^2)^{-(a+1+K/2)}}{(2\pi)^{K/2}|\boldsymbol{V}|^{1/2}\Gamma(a)} \exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}-\boldsymbol{m})^T\boldsymbol{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m}) + 2b\right)\right), \\ p(\eta) &= \frac{\Gamma(r+t)}{\Gamma(r)\Gamma(t)}\eta^{r-1}(1-\eta)^{t-1}I_{[0,1]}(\eta). \end{aligned}$$

By direct calculation it can be verified that

$$\begin{aligned} &(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{m})^T\boldsymbol{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m}) \\ &= \boldsymbol{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + \boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m} - (\boldsymbol{m}^*)^T(\boldsymbol{V}^*)^{-1}(\boldsymbol{m}^*) + \\ &\quad + (\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\boldsymbol{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{V}^* &= (\boldsymbol{V}^{-1} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\ \boldsymbol{m}^* &= \boldsymbol{V}^*(\boldsymbol{V}^{-1}\boldsymbol{m} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}). \end{aligned}$$

With this notation the joint density $P(\boldsymbol{Y}, \boldsymbol{\beta}, \sigma^2, \eta)$ is

$$\begin{aligned} &p(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, \eta)p(\boldsymbol{\beta}, \sigma^2, \eta) \\ &= p(\eta) \times \frac{b^a(\sigma^2)^{-(1+a+\frac{n+K}{2})}}{(2\pi)^{\frac{n+K}{2}}\Gamma(a)|\boldsymbol{V}|^{\frac{1}{2}}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \times \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}((\boldsymbol{\beta}-\boldsymbol{m}^*)^T(\boldsymbol{V}^*)^{-1}(\boldsymbol{\beta}-\boldsymbol{m}^*) + b^*)\right), \end{aligned}$$

where $b^* = 2b + \boldsymbol{Y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + \boldsymbol{m}^T\boldsymbol{V}^{-1}\boldsymbol{m} - (\boldsymbol{m}^*)^T(\boldsymbol{V}^*)^{-1}(\boldsymbol{m}^*)$. By noticing that as a function of $\boldsymbol{\beta}$ the above density is proportional to the density of $\mathcal{N}(\boldsymbol{m}^*, \sigma^2\boldsymbol{V}^*)$, we are able to integrate analytically with respect to $\boldsymbol{\beta}$:

$$p(\boldsymbol{Y}, \sigma^2, \eta) = \frac{p(\eta)b^a}{(2\pi)^{\frac{n}{2}}\Gamma(a)|\boldsymbol{V}|^{\frac{1}{2}}}\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}^*|}\right)^{-\frac{1}{2}}(\sigma^2)^{-(1+a+\frac{n}{2})}\exp\left(-\frac{b^*}{2\sigma^2}\right).$$

As a function of σ^2 the above function is proportional to the density of Inv-Gamma $\left(a+\frac{n}{2}, \frac{b^*}{2}\right)$ allowing us to calculate

$$p(\boldsymbol{Y}, \eta) = \frac{b^a\Gamma(a+\frac{n}{2})}{(2\pi)^{\frac{n}{2}}\Gamma(a)}\left(\frac{b^*}{2}\right)^{-(a+\frac{n}{2})}\left(\frac{|\boldsymbol{V}^*|}{|\boldsymbol{\Sigma}||\boldsymbol{V}|}\right)^{\frac{1}{2}}p(\eta). \tag{3.1}$$

As a function of η, the density (3.1) is proportional to the posterior of η, and thus evaluating it at a grid over the interval $[0, 1]$ allows us to do inference

on η and approximate the marginal likelihood of the data, $P(\boldsymbol{Y})$, by integrating (3.1) numerically. A Bayes factor, that is, the ratio of the marginal likelihoods of two models, can thus be calculated between models that differ in the structure of the predictor matrix $\boldsymbol{X}$ (e.g. testing genetic effects in GWAS), in the prior distributions of the parameters (e.g. whether $\eta = 0$), or both. Next we show how to do these computations efficiently by exploiting the same eigenvalue decomposition of $\boldsymbol{R}$ and transformed variables $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ that were introduced for maximum likelihood estimation in the previous section.

Consider the density (3.1), that is,

$$p(\boldsymbol{Y}, \eta) = c \times (b^*)^{-(a+\frac{n}{2})} \left(\frac{|\boldsymbol{V}^*|}{|\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} p(\eta), \text{ where } c = \frac{(2b)^a \Gamma(a+\frac{n}{2})}{\pi^{\frac{n}{2}} \Gamma(a) |\boldsymbol{V}|^{\frac{1}{2}}}$$

is independent of η. The goal is to integrate this over the interval $\eta \in [0, 1]$, for example, by evaluating it at a grid of m equally spaced points in $[0, 1]$ and by using the trapezoidal rule.

First we notice that $p(\eta)$ and $|\boldsymbol{\Sigma}|$ do not depend on $\boldsymbol{X}$ and thus we evaluate them once at the given grid points and store the results for repeated use with different $\boldsymbol{X}$ matrices. A similar idea is applied to the first three terms of the quantity

$$b^* = 2b + \boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y} + \boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{m} - (\boldsymbol{m}^*)^T (\boldsymbol{V}^*)^{-1} (\boldsymbol{m}^*).$$

The only $\boldsymbol{X}$ dependent quantities are thus

$$\begin{aligned} \boldsymbol{V}^* &= (\boldsymbol{V}^{-1} + \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \text{ and} \\ \boldsymbol{m}^* &= \boldsymbol{V}^* (\boldsymbol{V}^{-1} \boldsymbol{m} + \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}). \end{aligned}$$

Suppose that the analyzed $\boldsymbol{X}$ matrices differ from each other only in one predictor which is stored in the last column K of $\boldsymbol{X}$. Then only the element K of the vector $\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}$ needs to be recomputed:

$$(\boldsymbol{X}^T)_{K\bullet} \boldsymbol{\Sigma}^{-1} \boldsymbol{Y} = (\boldsymbol{X}^T)_{K\bullet} \boldsymbol{U} \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{U}^T \boldsymbol{Y} = (\widetilde{\boldsymbol{X}}^T)_{K\bullet} \widetilde{\boldsymbol{\Sigma}}^{-1} \widetilde{\boldsymbol{Y}} = \sum_{i=1}^{n} \frac{\widetilde{\boldsymbol{X}}_{iK} \widetilde{\boldsymbol{Y}}_i}{\eta(d_i - 1) + 1},$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}^T \boldsymbol{Y}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{U}^T \boldsymbol{X}$, and $\widetilde{\boldsymbol{\Sigma}} = \eta \boldsymbol{D} + (1 - \eta) \boldsymbol{I}$ as in the previous section and $(\boldsymbol{X}^T)_{K\bullet}$ denotes the row K of matrix $\boldsymbol{X}^T$. Similarly we can recompute the elements

$$(\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})_{Kj} = \sum_{i=1}^{n} \frac{\widetilde{\boldsymbol{X}}_{iK} \widetilde{\boldsymbol{X}}_{ij}}{\eta(d_i - 1) + 1}, \text{ for } j = 1, \ldots, K.$$

Since $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ have already been computed for the ML estimation by the conditional maximization algorithm, the additional complexity of these Bayesian computations is $\mathcal{O}(mKn)$ operations. (Here we assume that $K \ll n$ so that the complexity of the matrix operations on $K \times K$ matrices is negligible compared to the operations involving all n individuals.)

An existing Bayesian treatment of the linear mixed model considered uses parameterization with two variance components σ_ε^2 and σ_ϱ^2 which are related to our parameters as $\sigma_\varepsilon^2 = (1-\eta)\sigma^2$ and $\sigma_\varrho^2 = \eta\sigma^2$ (Sorensen and Gianola, 2002). When independent Inverse-Gamma priors are assigned to σ_ε^2 and σ_ϱ^2 it seems that it is not possible to analytically derive their one-dimensional marginal distributions (p. 323 Sorensen and Gianola (2002)). Thus, it seems that our parameterization has an advantage in computing marginal likelihoods since we only need to integrate numerically over the one-dimensional compact set $\eta \in [0,1]$ as opposed to the unbounded two-dimensional set $(\sigma_\varepsilon^2, \sigma_\varrho^2) \in (0,\infty) \times (0,\infty)$.

References.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11** 375-386.

Astle, W. and Balding, D. (2009). Population structure and cryptic relatedness in genetic association studies. *Stat Sci* **24** 451-471.

Denby, L., Kafadar, K. and Land, T. (1998). Modeling circuit boards yield. In *Statistical Case Studies: A Collaboration between Academe and Industry* (R. Peck, L. Haugh and A. Goodman, eds.) 143-150. ASA-SIAM.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, Baltimore, USA.

Kang, H., Zaitlen, N., Wade, C., Kirby, A., Heckerman, D., Daly, M. and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709-1723.

Kang, H., Sul, J., Service, S., Zaitlen, N., Kong, S., Freimer, N., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42** 348-354.

Nelder, J. and Wedderburn, W. (1972). Generalized Linear Models. *J R Statist Soc A* **135** 370-384.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics. Vol 2B. Bayesian Inference.*, 2nd ed. Arnold, London.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag. USA.

Matti Pirinen†, Peter Donnelly†◇, Chris Spencer†
†Wellcome Trust Centre for Human Genetics,
University of Oxford,
Roosevelt Drive
OX3 7BN,
Oxford, UK.
◇Department of Statistics,
University of Oxford,
1 South Parks Road
OX1 3TG,
Oxford, UK.
E-mail: matti.pirinen@iki.fi
E-mail: peter.donnelly@well.ox.ac.uk
E-mail: chris.spencer@well.ox.ac.uk